

Vision Language Models: A Survey of 26K Papers (CVPR, ICLR, NeurIPS 2023–2025)

Fengming Lin^{*1}

¹School of Computer Science, The University of Manchester, Manchester, UK

October 13, 2025

Abstract

We present a transparent, reproducible measurement of research trends across 26,104 accepted papers from CVPR, ICLR, and NeurIPS spanning 2023–2025. Titles and abstracts are normalized, phrase-protected, and matched against a hand-crafted lexicon to assign up to 35 topical labels and mine fine-grained cues about tasks, architectures, training regimes, objectives, datasets, and co-mentioned modalities. The analysis quantifies three macro shifts: (1) a sharp rise of multimodal vision–language–LLM work, which increasingly reframes classic perception as instruction following and multi-step reasoning; (2) steady expansion of generative methods—with diffusion research consolidating around controllability, distillation, and speed; and (3) resilient 3D and video activity, with composition moving from NeRFs to Gaussian splatting and a growing emphasis on human- and agent-centric understanding. Within VLMs, parameter-efficient adaptation (prompting, adapters/LoRA) and lightweight vision–language bridges dominate; training practice shifts from building encoders from scratch to instruction tuning and fine-tuning strong backbones; contrastive objectives recede relative to cross-entropy/ranking and distillation. Cross-venue comparisons show CVPR’s stronger 3D footprint and ICLR’s highest VLM share, while reliability themes (efficiency, robustness) diffuse across areas. We release the lexicon and methodology to enable auditing and extension. Limitations include lexicon recall and abstract-only scope, but the longitudinal signals are consistent across venues and years.

1 Introduction

The computer-vision and machine-learning community has undergone a visible transition in 2023–2025. With the consolidation of large-scale pretrained models (CLIP/BLIP/LLaVA [1, 2, 3] families, ViT backbones) and ubiquitous diffusion-based generators, the research focus has shifted: classic perception remains active, yet a large fraction of accepted papers are now organized around multimodal training, general-purpose reasoning or generation, and efficiency. While many reports provide anecdotal evidence, our goal is to quantify this transition from primary sources: the official titles and abstracts of accepted papers at CVPR, ICLR, and NeurIPS.

2 Data and Methodology

Data: We ingest all JSONL files collected by our Python spider: CVPR (2023: 2,353 papers; 2024: 2,713; 2025: 2,871), ICLR (2023: 4,372; 2024: 2,260; 2025: 3,704), and NeurIPS (2023: 3,337; 2024: 4,494). After removing empty records, 26,104 abstracts remain for analysis. In addition, for trend analysis only, we also include 8,424 papers from 2022; these 2022 records are used solely to study longitudinal trends and are excluded from content analysis, which focuses on the most recent three years.

Text processing: We normalize Unicode, lowercase, strip punctuation, protect multi-word phrases (e.g.,

^{*}Email: fengming.lin@manchester.ac.uk; sdulinfm@gmail.com

“gaussian splatting”, “neural radiance fields”, “vision language model”) as single tokens, and remove general stopwords and generic CV terms to retain the technical content.

Labeling: Each abstract is matched against 35 regular-expression categories (Diffusion, Vision-Language/LLM, 3D, Video, Robustness, Efficiency, etc.). A paper can receive multiple labels. We then compute prevalence as the fraction of abstracts in a year that match the label.

Fine-grained mining: For selected areas we search for sub-topics: tasks (e.g., grounding), architecture motifs (e.g., LoRA/adapters), training regimes (pretrain+fine-tune, instruction tuning), representative losses, and named datasets.

Caveat: The approach is transparent and replicable, but lexicon-driven—precision is high for canonical phrases; recall may miss niche synonyms; all numbers refer to abstracts only.

3 Macro Trends

Fig 1 summarizes the fraction of papers per year tagged as VLM/LLM, diffusion, 3D (NeRF/Gaussian + geometry), and video understanding et al. The rise of VLM is unmistakable: from 16% (2023) to 40% (2025) of all abstracts we analyzed. By 2025, the VLM share reaches 39.5% at CVPR and 40.7% at ICLR. Diffusion expands in parallel (8%→14.9%→19.2%), while classic 3D remains stable overall with composition shifting from NeRFs to Gaussian splatting. Video has a steady incline, partly due to video-LLMs and long-context modeling.

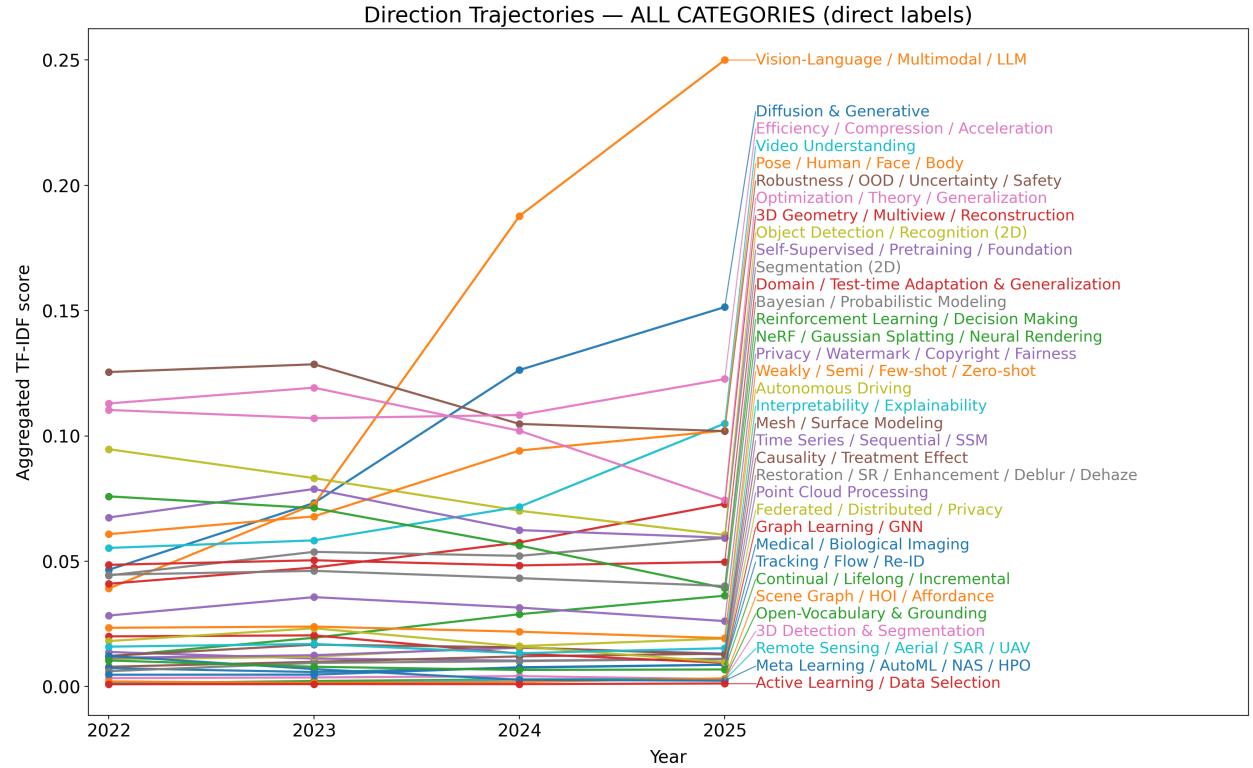


Figure 1: **Direction Trajectories across CVPR+ICLR+NeurIPS — ALL CATEGORIES (direct labels).** Each curve is the yearly aggregated TF-IDF mass for a direction (integer year ticks).

The consolidated view highlights three macro patterns: (i) a sharp takeoff of *VLM/LLM*; (ii) sustained growth of *Generative Model*, with increasing integration into perception pipelines; and (iii) steady increases in *Video Understanding* and stable but reconfigured *3D Reconstruction* trajectories.

Fig.2 presents small-multiple line charts in which each panel corresponds to a research direction. The horizontal axis shows years (2022–2025) and the vertical axis shows the fraction of all papers attributed to

that direction. Because each panel has its own vertical scale, comparisons should be made within a panel across years rather than across panels by absolute height. Two global trends stand out: (i) generative and multimodal areas expand steadily and spill over into 3D, video, and editing; and (ii) several classical learning paradigms (e.g., self-supervised, meta-learning, GNNs, weak supervision) decline or plateau in relative share, while “engineering and safety” themes such as efficiency, robustness, and privacy diffuse across the field.

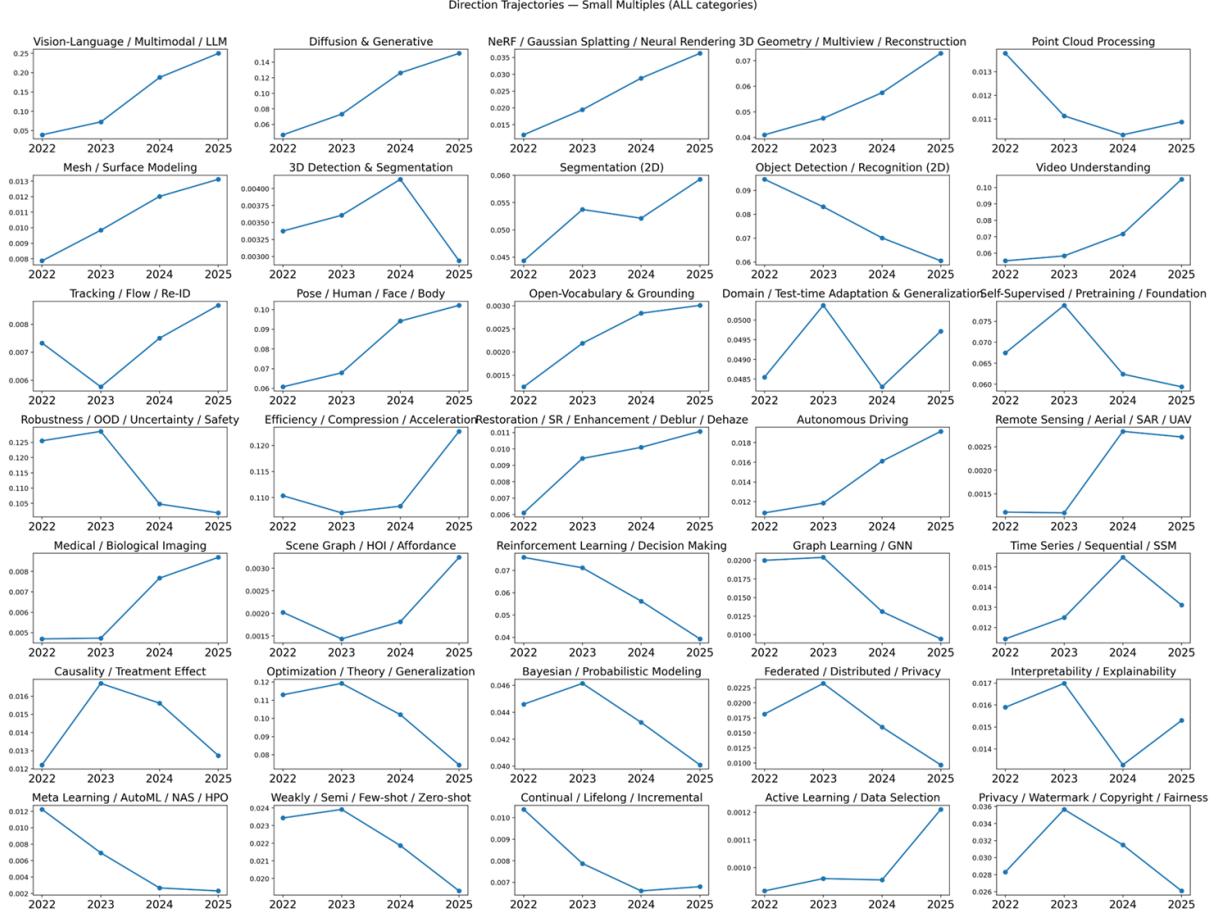


Figure 2: **Small-multiples view of research-direction trajectories from 2022–2025.** Each panel shows one category; the x-axis is year and the y-axis is normalized topic intensity (aggregated TF-IDF score). Vision-Language/Multimodal/LLM, Diffusion & Generative, and NeRF/Neural Rendering trend upward, while some traditional areas (e.g., 2D Object Detection, Self-Supervised/Pretraining) soften; others (e.g., GNN, Bayesian, Optimization) remain flat or decline slightly.

Generative and multimodal topics show the most pronounced rise. Diffusion and generative methods increase year over year, indicating that content generation continues to drive both methodological and application-level innovation. In parallel, vision-language and broader multimodal directions grow rapidly, reflecting the consolidation of large models and cross-modal alignment as core infrastructure. Closely related 3D content representations also climb: Gaussian splatting and neural rendering trend upward, as do 3D geometry, multiview, and reconstruction, signaling a shift from recognition alone toward high-fidelity synthesis and editable 3D assets.

Structure-aware 3D understanding strengthens as well. After an early dip, point-cloud processing rebounds slightly, while mesh and surface modeling rise steadily, suggesting interest in controllable, constraint-aware geometry. 3D detection and segmentation fluctuate: 3D detection peaks mid-period and then recedes, and 2D segmentation dips and recovers in the latest year. These oscillations likely reflect evolving benchmarks, task boundaries, and downstream deployment needs.

Temporal perception and human-centric understanding gain traction. Video understanding climbs from a low baseline, and tracking, flow, and re-identification increase gradually. Pose, face, and full-body analysis accelerate in the last two years, underscoring the move toward agent- and human-centered applications. Open-vocabulary grounding continues to rise, indicating that zero-shot recognition with language priors is becoming a standard capability.

Some “inner-loop” methodology themes cool or integrate into larger systems. Self-supervised pretraining peaks around 2023 and then declines, consistent with the field’s pivot to adapting foundation models rather than foregrounding self-supervision as a standalone contribution. Meta-learning and AutoML, weak/semi-/few-shot learning, GNNs, and causality/treatment-effect topics trend downward or remain choppy, suggesting these ideas increasingly appear as modules within broader pipelines rather than as primary focal points. Optimization and theory also edge downward, possibly due to attention shifting toward system-level integration and empirical capability building.

Engineering and reliability concerns become more visible. Efficiency, compression, and acceleration surge in the most recent year, while robustness, out-of-distribution generalization, uncertainty, and safety show steady growth. Privacy, watermarking, copyright, and fairness recede from a prior peak but maintain a notable presence, indicating that trust and governance have normalized rather than vanished. Interpretability rebounds after a dip, and federated and distributed learning stabilize after an earlier high, reflecting the ongoing importance of multi-institution collaboration and data-side constraints in real deployments.

Application-oriented areas display differentiated momentum. Medical and biological imaging rise consistently. Scene graphs, human-object interaction, and affordances strengthen recently, marking a shift from static recognition to interaction-ready understanding. Restoration, super-resolution, and enhancement track upward steadily; autonomous driving is broadly stable with a slight increase; and remote sensing maintains a modest uptick. Time-series, sequential modeling, and state-space approaches reach a high around 2024 and soften thereafter, consistent with their absorption into the foundation-model toolkit. Active learning and data selection pick up in the latest year, highlighting renewed attention to data efficiency and dataset governance.

Overall, Fig.2 depicts an ongoing transition toward “multimodal generative foundations plus 3D perception and editing,” while traditional paradigm-centric methods recede as independent flags and reappear as components inside larger systems. Simultaneously, scaling-aware and safety-oriented concerns grow in prominence, pushing research toward solutions that are efficient, robust, and compliant. For problem selection, the curves encourage emphasis on cross-modal, 3D, and video settings anchored in human-centric tasks, together with system designs that explicitly target efficiency, reliability, and data governance.

In Fig 3, across all venues, *Vision-Language/Multimodal/LLM* exhibits the steepest increase, followed by *Diffusion & Generative*. *Video Understanding* and human-centric topics rise as well, while 3D-related areas (*3D Geometry; NeRF/Gaussian Splatting*) continue to gain but at a slower rate.

4 Vision–Language and LLMs

Executive summary. Across CVPR/NeurIPS/ICLR (2023–2025), VLM abstracts pivot from *grounding/referring* toward *instruction following and reasoning*. Parameter-efficient adaptation (adapters/LoRA) and prompt-style mechanisms remain common. Training is dominated by *pretrain + finetune*, with a clear rise in *instruction tuning*. Loss design shifts away from purely contrastive objectives toward mixtures that include KL/distillation and cross-entropy/ranking. Explicit dataset name mentions become rarer in abstracts (COCO/ImageNet steadily decline). Co-mentioned modalities tilt toward 3D and depth, while audio stabilizes and begins to recover in 2025.

4.1 Models (named backbones & families)

ALIGN [4] remains the single most cited family in VLM abstracts ($\approx 5.8\%$ in 2024, dipping slightly to 5.1% in 2025). *LLaVA* [3] shows the fastest growth ($0.1\% \rightarrow 1.2\% \rightarrow 2.7\%$), mirroring the community’s shift to instruction-following VLMs. Classical backbones shrink in visibility—*ResNet/ConvNeXt* [5, 6] and *ViT* [7] roughly halve by 2025. *MoE* [8] references roughly double by 2025, indicating increasing interest in expert routing for multimodal scaling.

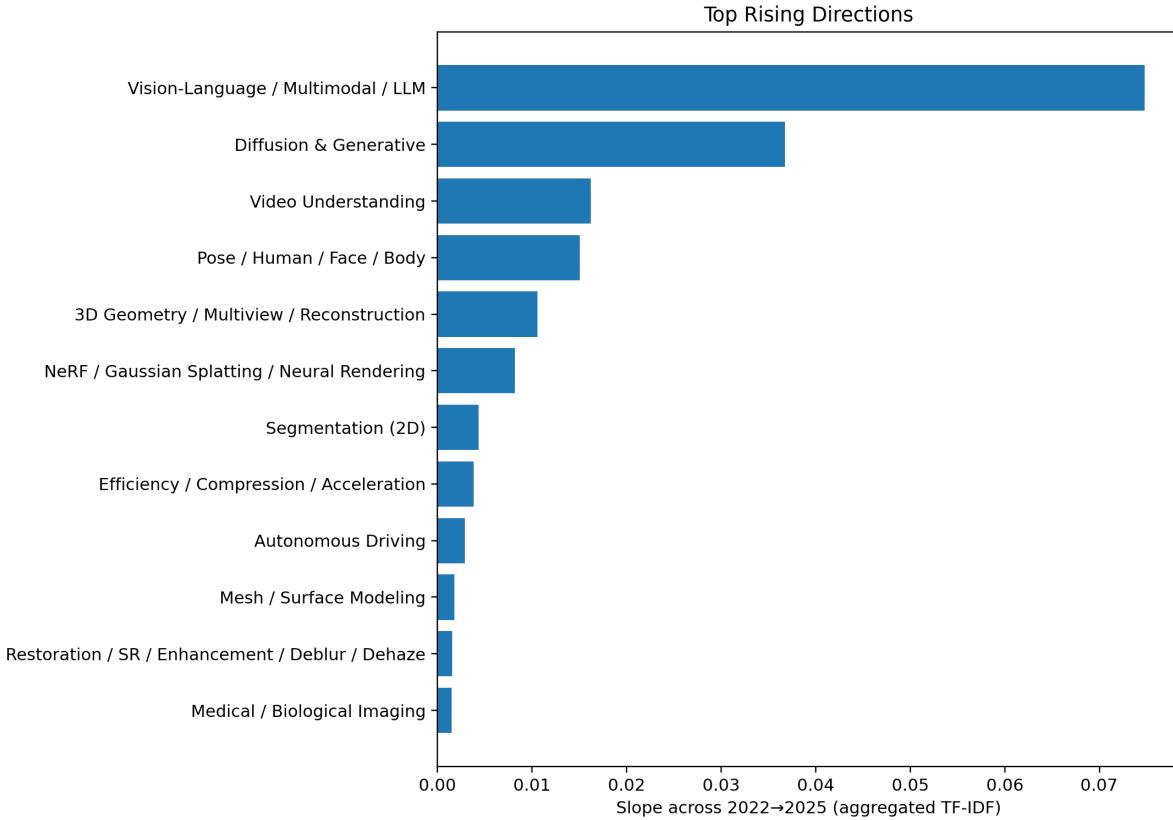


Figure 3: **Top Rising Directions across CVPR+ICLR+NeurIPS (2022–2025).** Bars show the slope of each direction’s aggregated TF-IDF trajectory over years. Larger slope = faster growth.

ALIGN [4] (“A Large-scale ImaGe and Noisy-text embedding”) scales contrastive dual-encoder pretraining to over one billion noisy image–alt-text pairs, deliberately avoiding heavy curation and showing that scale can compensate for label noise. A vision encoder and a text encoder are trained with a contrastive objective to align representations in a shared space, enabling strong *zero-shot classification* and *cross-modal retrieval* without task-specific heads. The paper emphasizes that simple frequency filtering plus large scale suffices to surpass more complex cross-attention models on Flickr30k/MSCOCO retrieval.

CLIP [1] learns image–text alignment using a dual-encoder (image transformer/CNN + text transformer) trained with InfoNCE-style contrastive loss on hundreds of millions of web pairs, producing a universal embedding space for *zero-shot recognition* and robust transfer. CLIP commonly serves as the frozen vision encoder in later LVLMs. (For scale context, BLIP reports the CLIP baseline trained on ~400M pairs in its retrieval comparison table.)

BLIP [2] (“Bootstrapping Language–Image Pretraining”) introduces a Multimodal Mixture of Encoder–Decoder (MED) that can act as (*i*) unimodal encoders for ITC, (*ii*) an image-grounded text encoder for ITM, and (*iii*) an image-grounded text decoder for language modeling—jointly optimizing ITC, ITM, and LM to support both understanding and generation. BLIP also proposes CapFilt (captioning & filtering) to *bootstrap* cleaner pretraining data from noisy web alt-text, improving retrieval, captioning, and VQA across 14M–129M images and scaling to partially include LAION.

Flamingo [12] is an LVLM that conditions a large decoder-only LLM on visual evidence via gated cross-attention layers. A lightweight Perceiver-Resampler compresses image/video features into a small set of tokens fed into the language model, enabling *few-shot* multimodal learning on interleaved image–text streams. The work highlights training on a mixture of web corpora with interleaved modalities (e.g., *M3W*) to yield strong few-shot and zero-shot VQA/retrieval.

Table 1: Top models referenced by VLM papers (share of VLM abstracts).

Item	2023	2024	2025	Trend	Slope (pp/yr)
ALIGN [4]	4.3%	5.8%	5.1%	-0.8%	0.65
LLaVA [3]	0.1%	1.2%	2.7%	+2.6%	0.91
ResNet/ConvNeXt/CNN [6, 5]	2.9%	0.4%	0.5%	-2.4%	-0.74
ViT [7]	1.5%	1.2%	0.6%	-0.9%	-0.13
MoE [8]	0.6%	0.6%	1.3%	+0.6%	0.26
BLIP-2 [9]	0.6%	0.2%	0.2%	-0.4%	0.02
BLIP [2]	0.4%	0.1%	0.1%	-0.3%	-0.00
CLIP/OpenCLIP [1]	0.1%	0.3%	0.2%	+0.1%	0.06
GLIP [10]	0.4%	0.1%	0.0%	-0.3%	-0.06
Swin [11]	0.3%	0.1%	0.1%	-0.2%	-0.23
Flamingo [12]	0.2%	0.1%	0.0%	-0.2%	-0.01
GroundingDINO [13]	0.0%	0.2%	0.2%	+0.2%	0.06

LLaVA [3] (“Large Language-and-Vision Assistant”) shows that visual instruction tuning—supervising an LLM with *high-quality, GPT-4 generated multimodal conversations* paired with images—can endow the model with broad *instruction following*, reasoning, and perception skills. A simple projection bridge maps image features from a CLIP-like encoder into the LLM token space, after which stage-wise tuning on the synthetic conversations yields strong zero-shot generalization.

DINO [34] demonstrates that self-distillation without labels (student–teacher with momentum) applied to Vision Transformers yields emergent patch-level correspondences and high-quality features for downstream tasks; key ingredients are multi-crop augmentations and centering strategies that stabilize training. DINO is often used as an initialization or backbone in vision systems. DINOv2 [35] scales this recipe substantially: it trains larger ViT backbones on a carefully curated, de-duplicated image corpus, strengthens the teacher–student alignment with improved regularization and prototype objectives, and targets resolution-robust, task-agnostic features. As a result, DINOv2 delivers strong zero-/few-shot transfer and competitive dense recognition performance without labels, making it a common visual encoder for VLMs and robotics pipelines. DINOv3 [36] further advances self-supervised ViTs by (i) scaling both data and model size with careful data preparation/optimization, (ii) introducing Gram anchoring to prevent degradation of dense feature maps under long schedules, and (iii) applying post-hoc strategies that improve flexibility across input resolution, model size, and even text-alignment. The released DINOv3 suite yields high-quality dense features and outperforms prior self-/weakly-supervised visual foundation models across a broad range of tasks.

Grounding DINO [13] marries DINO-style detectors with open-vocabulary pretraining, aligning region-level visual features with free-form text so that *one detector* handles both traditional object categories and arbitrary phrases (e.g., referring expressions) in a single pipeline. The method leverages region–text pretraining and phrase grounding signals to enable *open-set* detection.

MoE [8] architectures (e.g., Switch Transformers) insert a sparse gating mechanism that *routes each token* to a small subset of specialized *expert* MLPs, achieving parameter-efficient scaling to hundreds of billions or trillion-parameter regimes while keeping per-token compute roughly constant. Such sparse decoders increasingly serve as the language backbone in LVLMs.

Overall, across these families, two design templates dominate: dual-encoder contrastive pretraining (ALIGN/CLIP) for universal embeddings, and LLM-centric conditioning with cross-attention and instruction tuning (Flamingo/LLaVA). BLIP bridges understanding–generation with its MED and *data bootstrapping*, while DINO and Grounding-DINO provide strong vision backbones and open-vocabulary grounding that LVLMs can exploit.

In Table 2, **Dual-Encoder** families (ALIGN/CLIP) emphasize **dataset scale** and a **contrastive** recipe for broad transfer. **LVLM** families (Flamingo/LLaVA) inject vision into a powerful language decoder via

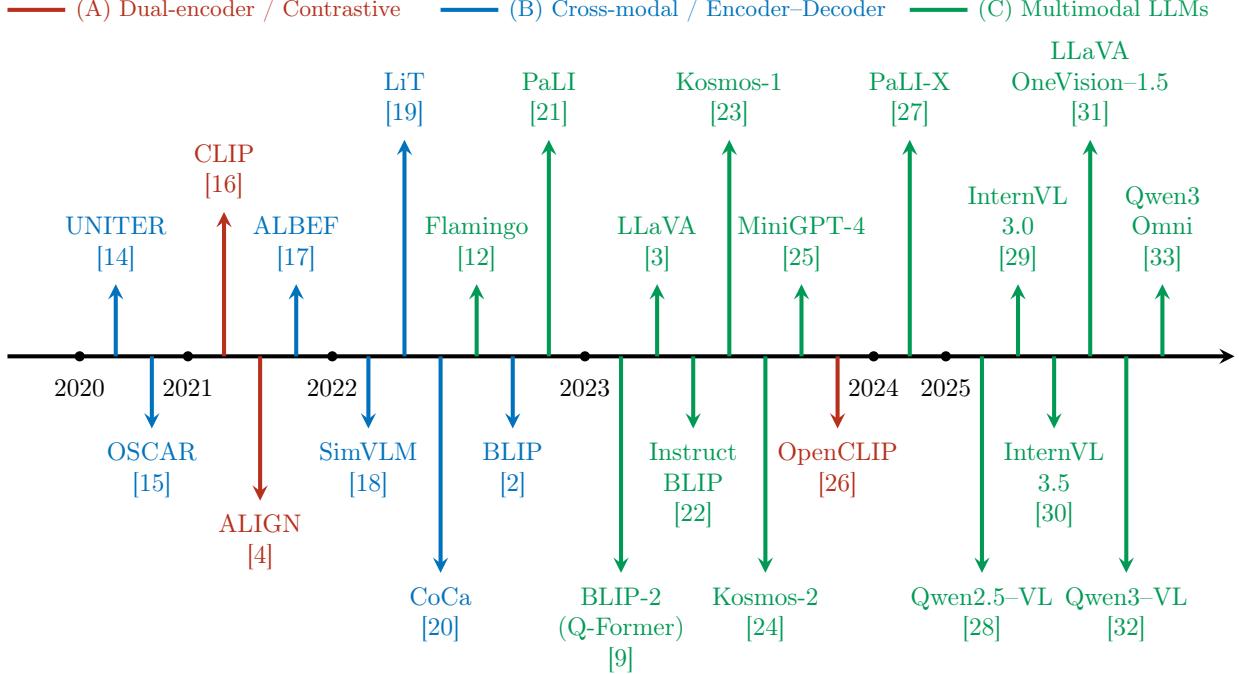


Figure 4: **Chronological overview of representative VLM / Multimodal LLM milestones (2020–2025).** Four color-coded categories: (A) Dual-encoder / Contrastive, (B) Cross-modal / Encoder-Decoder, (C) Multimodal LLMs. Each arrow is vertically offset within its year and labeled above/below with the model name and citation.

cross-attention or a simple **projector**, and then rely on **instruction-like supervision** to elicit reasoning/grounding. **BLIP** occupies the middle ground by unifying **ITC/ITM/LM** and **bootstrapping** cleaner captions. **DINO** and **Grounding DINO** secure the vision side—self-supervised features and open-vocabulary localization—that downstream LVLMs frequently build upon. Finally, **MoE** provides a path to scale **language backbones** with sparse compute, increasingly common in recent LVLMs.

4.2 Fusion / Architectural Integration

Table 3 summarizes how VLM papers integrate vision and language components. We observe three clear movements.

Lightweight promptization rises. Prompt/Prefix tuning is the most frequently referenced mechanism and continues to trend upward ($13.0 \rightarrow 16.4 \rightarrow 14.3\%$, Trend $+1.3, pp$; Slope $+3.43, pp/yr$), reflecting a preference for parameter-efficient adaptation on strong frozen backbones [37, 38]. Adapter/LoRA usage also grows steadily ($1.3 \rightarrow 4.0 \rightarrow 4.1\%$, Trend $+2.8, pp$), consistent with the widespread adoption of low-rank updates for multimodal fine-tuning [39].

Bridging modules remain active while “heavy” fusion softens. Cross-/co-attention stays stable to slightly positive (Trend $+0.5, pp$), often appearing in systems that inject visual features into frozen LMs (e.g., Flamingo-style cross-attention layers) [12]. Projector/MLP heads increase (Trend $+0.6, pp$), consistent with simple alignment layers that map modality embeddings into a shared space (as in CLIP-style dual encoders with learned projections) [1]. The Q-Former bridge (BLIP-2) stays non-negligible and flat (Trend $+0.0, pp$), suggesting continued but specialized use where learned query tokens are advantageous [9].

Architecture choices rebalance. Mixture-of-Experts/gating is mildly up (Trend $+0.8, pp$), indicating exploratory interest in sparse capacity [40]. By contrast, encoder-decoder patterns show a small net decline (Trend $-1.3, pp$), despite remaining important for generation-centric systems (e.g., OFA, PaLI) [41, 21]. Dual-encoder/two-tower usage edges downward (Trend $-0.1, pp$), likely reflecting a shift from pure retrieval/contrastive training toward instruction-following and generative pipelines where cross-attention or

Table 2: **Named model families and design choices.** We summarize core ideas, fusion styles, typical backbones, objectives, and the nature/scale of pretraining data when explicitly reported in the cited papers. Abbreviations: ITC/ITM = image–text contrast / matching; LM = language modeling; GCA = gated cross-attention.

Model	Core Idea	Fusion	Vision & Text Backbones	Objective(s)	Pretraining Data
ALIGN [4]	Dual-encoder at web scale	None (late)	CNN/Transformer encoders	Contrastive alignment	~1B+ noisy alt-text pairs
CLIP [1]	Dual-encoder, universal embeddings	None (late)	ViT/ResNet + text Transformer	Contrastive (InfoNCE)	~400M web pairs
BLIP [2]	MED unifies understanding & generation + CapFilt	Cross-attn in en-coder/decoder	ViT + BERT-style text Transformer	ITC + ITM + LM	14M–129M images (+ partial LAION)
Flamingo [12]	LVLM with GCA layers and Perceiver Resampler	Gated cross-attention	CNN/ViT features → LLM decoder	Autoregressive LM (vision-conditioned)	Interleaved image–text corpora (e.g., M3W)
LLaVA [3]	Visual instruction tuning with GPT-4 conversations	Projector → LLM tokens	CLIP-like image encoder + decoder LLM	SFT / LM on multimodal dialogs	High-quality synthetic conversations
DINO [34]	Self-distillation ViT without labels	N/A	Vision Transformer (student/teacher)	Self-distill with momentum teacher	Unlabeled images (multi-crop)
Grounding DINO [13]	Open-vocabulary detection by region–text pretraining	Cross-modal at region level	Detector + text encoder (phrase grounding)	Detection + grounding losses	Region–phrase corpora (web & annotations)
MoE [8]	Token routing to expert MLPs (sparse)	Inside decoder (sparse)	Decoder-only Transformer with experts	Autoregressive LM, sparse gating	Large text corpora (scalable)

bridging modules provide tighter fusion.

Overall, the field is converging on parameter-efficient tuning + light bridging as default design knobs (for cost, stability, and reusability), while reserving cross-attention/Q-Former for cases that require stronger, token-level conditioning. Sparse MoE appears as an emerging capacity-scaling option, whereas monolithic encoder–decoder and pure dual-encoder designs, though still influential, grow more selectively applied.

4.3 Tasks

Table 4 profiles how task emphases have evolved in recent VLM work (shares over 2023–2025). We see a decisive pivot toward instruction-following and multi-step reasoning, alongside a broad cooling of classic grounding/captioning style tasks.

Reasoning / Instruction: The fastest-growing stratum rises from 13.5% to 25.0% (Trend +11.5,pp; Slope +5.71,pp/yr), driven by instruction-tuned LMMs that attach visual adapters to frozen (or lightly fine-tuned) LMs and are optimized with conversation-style data (e.g., LLaVA, InstructBLIP, MiniGPT-4) [3, 22, 25]. These systems prioritize chain-of-thought localization, tool-use hooks, and preference optimization, naturally expanding the proportion of “reasoning” papers.

Grounding / Referring: Once dominant, this stratum declines by 13.0,pp overall, reflecting a shift from RefCOCO/phrase-grounding formulations toward instruction-following stacks where grounding appears as a

Table 3: **Fusion/architectural integration mechanisms within VLM papers.** “Trend” is the three-year net change (2025 minus 2023, in percentage points). “Slope” is the least-squares linear slope across 2023–2025.

Item	2023	2024	2025	Trend	Slope (pp/yr)
Prompt/Prefix Tuning	13.0%	16.4%	14.3%	+1.3%	3.43
Adapter/LoRA	1.3%	4.0%	4.1%	+2.8%	1.26
Cross-/Co-attention	1.7%	2.2%	2.2%	+0.5%	0.46
Projector/MLP Head	0.9%	1.2%	1.5%	+0.6%	0.21
MoE/Gating	0.6%	0.6%	1.4%	+0.8%	0.24
Encoder-Decoder	1.6%	0.7%	0.3%	-1.3%	-0.39
Dual-encoder/Two-tower	0.3%	0.2%	0.1%	-0.1%	-0.04
Q-Former Bridge	0.0%	0.1%	0.0%	+0.0%	0.03

Table 4: Task strata within VLM papers (share).

Item	2023	2024	2025	Trend	Slope (pp/yr)
Reasoning/Instruction	13.5%	22.3%	25.0%	+11.5%	5.71
Grounding/Referring	25.9%	14.5%	12.9%	-13.0%	-8.36
Retrieval	8.5%	6.8%	8.3%	-0.2%	0.53
Captioning	6.2%	4.9%	4.4%	-1.9%	-0.53
VQA	2.4%	2.0%	1.9%	-0.5%	-0.05
Video QA/Captioning	1.3%	1.0%	1.2%	-0.1%	0.02
OCR/Text Recognition	0.9%	1.0%	1.0%	+0.2%	-0.44
Open-Vocabulary Det./Seg.	1.4%	0.8%	0.4%	-1.0%	-0.12

sub-capability rather than the end task. Methodologically, open-vocabulary detectors (e.g., GLIP, GroundingDINO) remain key components but are now more often embedded as plug-ins [10, 13].

Retrieval: The share is broadly stable (8.5,→,8.3%), consistent with the continued role of dual-encoders for retrieval and RAG back-ends (CLIP-style alignment; BLIP/OFA hybrids for generative read-out) [1, 2, 41].

Captioning and VQA: Generic image captioning declines modestly (−1.9,pp), likely absorbed by instruction-chat settings where captioning becomes an intermediate capability [2]. VQA dips slightly (−0.5,pp), as community benchmarks (e.g., VQA v2, GQA) give way to broader multimodal instruction suites; nonetheless VQA remains a standard probe for perception-reasoning coupling [42, 43].

Video QA / Captioning: The slice is nearly flat, indicating steady but not explosive growth. While datasets such as MSRVTT-QA and NExT-QA continue to anchor evaluation, compute/data cost keeps video tasks a smaller share relative to image-centric reasoning [44, 45].

OCR / Text recognition: The share is small but persistent (roughly 1%), reflecting increased demand for document-centric VLMs (e.g., TrOCR/Donut lines) and the prevalence of text-rich scenes within instruction datasets [46, 47].

Open-vocabulary detection / segmentation: A gradual decline (−1.0,pp) likely reflects consolidation: many recent LMMs use open-vocab perception as a front-end (OWL-ViT/ViLD/SAM/OpenSeg families) rather than publishing it as a standalone task [48, 49, 50, 51].

Overall, the center of gravity is moving from task-specific supervision (grounding, captioning) toward generalist, instruction-tuned reasoning with retrieval and open-vocab perception as composable sub-systems. This aligns with the design trend noted in Sec. 4.2: lightweight adaptation of powerful frozen LMs plus modular visual front-ends.

4.4 Training Paradigms

Table 5 tracks the training regimes emphasized in recent VLM papers (2023–2025). We observe a decisive shift toward *fine-tuning strong pretrained components with parameter-efficient knobs and instruction data*, while earlier web-scale weak/ self-supervision plays a relatively smaller role in new papers.

Table 5: Training paradigms appearing in VLM papers.

Item	2023	2024	2025	Trend	Slope (pp/yr)
Pretrain + Finetune	11.6%	16.9%	16.8%	+5.2%	3.83
Prompt/Prefix	13.0%	16.4%	14.3%	+1.3%	3.43
Self-/Weak-/Semi-sup.	9.6%	2.8%	3.5%	-6.1%	-2.65
Distillation	4.2%	4.8%	4.0%	-0.4%	0.56
Instruction Tuning	1.1%	4.2%	5.0%	+3.9%	1.75
LoRA/Adapters	1.3%	4.0%	4.1%	+2.8%	1.26
Multi-task/Curriculum	2.6%	1.6%	1.9%	-0.8%	-0.08

Pretrain + Finetune: This remains the anchor recipe and grows by +5.2 pp overall. Typical patterns couple a frozen or lightly finetuned LM with an image encoder that was pretrained via contrastive/web supervision (e.g., CLIP/ALIGN), then finetune end-to-end or at the bridge for downstream tasks [1, 4, 2].

Prompt / Prefix: Promptization rises (+1.3 pp), reflecting the appeal of *tuning a small prompt vector* instead of full weights for new domains or tasks [37, 38]. In multimodal stacks this often means learning visual prompts or cross-modal prompts while keeping the LM mostly frozen.

Instruction Tuning: A substantial increase (+3.9 pp) mirrors the community pivot to conversational VLMs: systems such as LLaVA and InstructBLIP attach vision adapters and then perform supervised instruction tuning on (often synthetic) multimodal dialogues [3, 22]. This training style turns captioning/grounding capabilities into general instruction following.

LoRA / Adapters: Parameter-efficient updates expand (+2.8 pp), driven by their low compute/VRAM cost and stability when merging or stacking skills across domains [39, 52]. In practice, many VLMs combine LoRA with instruction tuning.

Self- / Weak- / Semi-supervision: The share decreases (−6.1 pp). While self/weak supervision (e.g., SimCLR, MoCo; large-scale noisy web pairs as in ALIGN/CLIP) remains crucial for *pretraining* encoders, new papers increasingly take such backbones as given and focus on finetuning/instruction stages instead [53, 54, 1, 4, 55].

Distillation: Usage is relatively steady (net −0.4 pp; slight positive slope), often appearing to compress vision encoders, align modality bridges, or transfer reasoning supervision from stronger teachers into smaller student LMMs [56, 2].

Multi-task / Curriculum: A small decline (−0.8 pp) likely reflects consolidation: unified instruction tuning increasingly subsumes the benefits of multi-task/curriculum training observed in earlier generalist frameworks such as OFA/PaLI and instruction-style curricula (FLAN) [41, 21, 57].

Overall, the center of gravity moves from building encoders with massive weak supervision to *adapting* those encoders (and frozen LMs) with instruction data and parameter-efficient updates. This reduces cost, speeds iteration, and aligns with the modular fusion trends in Sec. 4.2.

4.5 Loss Families

Table 6 summarizes the loss objectives most commonly reported in VLM papers from 2023–2025. We observe a *rebalancing from pretraining-style contrastive objectives toward alignment, supervision, and compression losses* that better fit instruction-tuned, generation-oriented pipelines.

Contrastive / InfoNCE: The share of contrastive learning drops markedly (−5.7 pp; slope −2.07 pp/yr). This mirrors a shift in new works away from building image–text encoders from scratch (as in CLIP/ALIGN) toward *adapting* such pretrained encoders and coupling them with large LMs. Contrastive learning remains

Table 6: Loss families in VLM papers.

Item	2023	2024	2025	Trend	Slope (pp/yr)
Contrastive/InfoNCE	10.8%	5.6%	5.1%	-5.7%	-2.07
KL/Distillation	5.6%	6.6%	5.8%	+0.3%	0.78
Triplet/Ranking	1.0%	0.7%	0.5%	-0.5%	-0.00
Cross-Entropy/Focal	0.8%	0.3%	0.6%	-0.1%	-0.19
MSE/L1/L2	0.3%	0.4%	0.3%	-0.0%	-0.10
Dice/IoU	0.4%	0.2%	0.1%	-0.3%	-0.06
Chamfer/EMD	0.1%	0.2%	0.0%	-0.1%	-0.10

foundational for representation learning and retrieval [58, 53, 1], but appears less frequently as the *primary* objective in papers that focus on multimodal instruction following.

KL / Distillation: A small net increase (+0.3 pp) reflects steady use of Kullback–Leibler-based distillation to compress vision backbones, align vision–language bridges, or transfer reasoning signals from stronger LMM teachers to lighter students [56, 2]. Distillation is often combined with supervised instruction losses to stabilize training.

Triplet / Ranking: This family remains low and stable. Triplet and margin-based ranking continue to appear in retrieval and grounding submodules, but are less central when the end task is dialogue-style generation [59].

Cross-Entropy / Focal: Classification-style objectives are modestly represented. They are widely used for detector heads or token-level supervision in bridging modules; focal loss is occasionally adopted for long-tail/open-vocabulary settings [60].

MSE / L1 / L2: Regression losses persist at a low but steady rate, mainly for projector heads, alignment regressors, or value heads in reward/prefence optimization stages. Their prevalence is limited by the dominance of token-level cross-entropy in instruction tuning.

Dice / IoU: These segmentation-oriented objectives shrink slightly as standalone open-vocabulary perception becomes a plug-in capability, rather than a primary contribution in VLM papers [61, 62].

Chamfer / EMD: Point-set distances (Chamfer, Earth Mover’s) are niche in mainstream VLM work and mostly appear in papers that also target 3D grounding or generative geometry, where set/shape supervision is required [63].

Overall, compared to 2023, losses in 2025 VLM papers reflect a community focus on *instruction-tuned generation with modular perception*: contrastive losses are still crucial for encoders and retrieval, but many new contributions emphasize KL-based distillation and CE-style supervision layered on top of pretrained backbones and LMs.

4.6 Datasets

Table 7 lists curated datasets that authors explicitly mention in VLM abstracts. We caution that abstracts under-report training data, especially for recent generalist LMMs that rely on mixtures of private or filtered corpora. Even with that caveat, three patterns emerge. First, legacy caption benchmarks such as MS-COCO [64, 65] and ImageNet [66] decline steadily in mentions (COCO: -3.0 pp; ImageNet: -1.5 pp). They remain common for sanity checks and zero-shot probes, but fewer new papers position them as primary contributions. Second, the share of open-web sources is small but persistent: LAION [67, 68] stays roughly flat overall, consistent with a trend where authors reference large web datasets generically while focusing the abstract on downstream instruction tuning. Third, task-specific datasets shrink or stabilize, reflecting a shift toward instruction-chat evaluations where many traditional tasks appear as sub-skills rather than endpoints. RefCOCO/RefCOCO+/RefCOCOg for referring expressions [69, 70], Flickr30k for captioning [71], CC3M/CC12M for web-scale caption pretraining [72, 73], VQA-v2 and OK-VQA for QA and knowledge-augmented QA [42, 74], WebVid/MSR-VTT/MSVD for video-text pairs [75, 76, 77], YouCook2/HowTo100M for instructional video grounding [78, 79], and Visual Genome for dense region-language grounding [80] all show flat to negative trends in abstract mentions.

Table 7: Curated datasets explicitly named in VLM abstracts (note under-reporting).

Item	2023	2024	2025	Trend	Slope (pp/yr)
MS-COCO	4.9%	2.1%	1.0%	-3.0%	-1.50
ImageNet	3.1%	2.4%	1.6%	-1.5%	-0.76
LAION	0.6%	0.8%	0.2%	-0.5%	0.03
RefCOCO/g/+	0.6%	0.6%	0.3%	-0.3%	-0.12
Flickr30k	0.8%	0.3%	0.2%	-0.7%	-0.28
CC3M/CC12M	0.4%	0.3%	0.3%	-0.1%	-0.19
VQA-v2/OK-VQA	0.4%	0.2%	0.3%	-0.2%	-0.09
WebVid/MSRVTT/MSVD	0.3%	0.3%	0.1%	-0.2%	-0.04
YouCook2/HowTo	0.2%	0.2%	0.1%	-0.1%	-0.16
Visual Genome	0.3%	0.1%	0.0%	-0.2%	-0.05
COCO Captions	0.1%	0.0%	0.0%	-0.0%	-0.03

Overall, abstracts increasingly de-emphasize named, single-dataset training and emphasize model behavior on multi-task, instruction-style suites. This aligns with the broader movement in Sections 4.2 and 4.4 toward reusing strong pretrained encoders, lightweight adaptation, and instruction tuning, with classic datasets retained primarily for probing, ablations, and comparability.

4.7 Modalities (co-mentioned with VLM)

Table 8 tracks which *additional* modalities are co-mentioned alongside vision–language modeling in 2023–2025 abstracts. Three observations emerge. First, 3D/point–cloud signals tick upward overall (Trend +0.7 pp), echoing growing interest in grounding LMMs to embodied/3D settings (robotics, digital twins, scene understanding). Much of this activity aligns features across modalities rather than building fully new architectures, e.g., shared embedding spaces that can accept RGB, depth, audio, IMU, or 3D point inputs as in ImageBind’s multi-sense alignment [81]. Second, classic image–text mentions decline (−3.5 pp), consistent with a shift from single-pair supervision toward broader instruction-style training where images appear interleaved with other cues (video frames, depth maps, layout, etc.). Depth/RGB-D is roughly steady to slightly positive (−0.2 pp net but a small positive slope), reflecting practical pipelines that fuse RGB with monocular or sensor depth for perception and 3D reasoning, again typically through lightweight adapters or shared-token bridges rather than bespoke RGB-D encoders. Third, audio/speech and video–text show modest declines (−1.4 pp and −1.1 pp, respectively). These modalities remain active, but many new papers focus on generalist instruction-following while citing prior audio/video-capable foundations instead of proposing stand-alone audio- or video-specific objectives; e.g., Flamingo-style cross-attention handles interleaved frames and extends naturally to audio features when available [12].

Table 8: Additional modalities co-mentioned in VLM papers.

Item	2023	2024	2025	Trend	Slope (pp/yr)
3D/Point Cloud	10.7%	11.7%	11.4%	+0.7%	0.71
Image-Text	8.6%	6.2%	5.1%	-3.5%	-0.42
Depth/RGB-D	4.7%	4.1%	4.6%	-0.2%	0.22
Audio/Speech	5.4%	3.5%	4.0%	-1.4%	-0.99
Video-Text	2.7%	1.5%	1.6%	-1.1%	-0.31

Overall, the co-mention trends suggest a pragmatic strategy: reuse strong image–text roots and *attach* additional modalities via alignment or prompting, reserving heavy bespoke modeling only where task benefits

are clear (e.g., robotics 3D grounding or long-horizon video understanding). This is consistent with the broader design and training shifts reported in Sections 4.2 and 4.4.

Takeaways for readers about VLMs. (1) If your problem can be framed as *instruction* or *reasoning*, emphasize it in the title/abstract to align with the strongest tailwinds. (2) Parameter-efficient *adapters*/LoRA and prompt-style interfaces have become lingua franca—treat them as strong baselines or ablations. (3) When contrastive pretraining is not the core contribution, keep it minimal and focus on downstream instruction/finetuning strategy and broad evaluation. (4) Explicit dataset name-dropping is rarely needed unless the dataset itself is a contribution; for video/3D, clarifying scale and diversity is more informative. (5) If you use 3D/depth/audio, state it early; cross-modal grounding is increasingly valued.

5 Cross-venue Comparison and Practical Advice

Cross-venue comparison. CVPR retains the strongest 3D emphasis (e.g., 23.1% of 2025 CVPR abstracts mention 3D geometry vs. 7.8% at ICLR), while ICLR has the largest 2025 VLM share (40.7%, comparable to CVPR’s 39.5%). CVPR sees the highest diffusion share in 2025 (25.7%) among the three. NeurIPS—available up to 2024 in our data—shows an early VLM ramp (30.5% in 2024) with diffusion at 11.6%.

Actionable advice for readers. (1) *Connect perception with VLMs.* Successful 2025 papers increasingly formulate classic vision problems (detection, segmentation, tracking) as instruction-following, grounding, or tool-using tasks on top of pretrained multimodal backbones. Designing lean projectors/adapters and strong data recipes (instruction or preference alignment) is impactful. (2) *Make generative tools usable.* If diffusion is part of the pipeline, emphasize controllability, speed/distillation, and reliability; benchmarks with precise quality/latency trade-offs resonate with current trends. (3) *Think long-context and video.* Methods that scale to minute- or hour-long sequences while preserving reasoning ability and memory efficiency are rising. (4) *Be explicit about efficiency and safety.* Lightweight inference, sparsity, cache-aware design, and safety/robustness concerns are common acceptance signals across venues.

6 Limitations

Our pipeline is lexicon-driven on abstracts only. Some fields (e.g., datasets, losses) are systematically under-reported in abstracts; therefore the absolute numbers are conservative. Papers may be multi-label; percentages are fractions of all papers per year, not summing to 100%. Nonetheless, the main trends (VLM ascent, diffusion pragmatization, steady 3D, rising video) are robust across venues and years.

7 Conclusion

We provide a transparent, replicable measurement of what the community has worked on in 2023–2025. By 2025, multimodal VLMs have become the organizing center for a large portion of accepted papers; diffusion matured into controllable, accelerated modules; and 3D and video remain vibrant with more cross-modal formulations. We release the full lexicon and code (in the conversation history) to encourage reproducibility and extension to other venues/years.

References

- [1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.
- [2] Junnan Li, Dongxu Li, Jianfeng Xie, and Steven C.H. Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *NeurIPS*, 2022.
- [3] Haotian Liu et al. LLaVA: Large language and vision assistant. *arXiv preprint arXiv:2304.08485*, 2023.

- [4] Chao Jia, Yinfei Yang, Ye Xia, et al. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning (ICML)*, 2021.
- [5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [6] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.
- [7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [8] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.
- [9] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023.
- [10] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10965–10975, 2022.
- [11] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.
- [12] Jean-Baptiste Alayrac et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022.
- [13] Shilong Liu, Zhao.yang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer, 2024.
- [14] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: Universal image-text representation learning. In *European Conference on Computer Vision (ECCV)*, 2020.
- [15] Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Lei Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. OSCAR: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision (ECCV)*, 2020.
- [16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.
- [17] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Gotmare, et al. ALBEF: Align before fuse for multi-modal learning. In *NeurIPS*, 2021.
- [18] Zirui Wang, Jianwei Bao, Alexey Bochkovskiy, et al. Simvlm: Simple visual language model pretraining with weak supervision. In *International Conference on Learning Representations (ICLR)*, 2022.
- [19] Xiaohua Zhai, Xiao Wang, Basil Mustafa, et al. Lit: Zero-shot transfer with locked-image text tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

- [20] Jiahui Yu et al. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.
- [21] Xi Chen, Xiao Wang, Saining Xie, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022.
- [22] Wenliang Dai et al. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023.
- [23] Siyuan Huang et al. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023.
- [24] Zhenyu Peng et al. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.
- [25] Deyao Zhu et al. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.
- [26] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, et al. Openclip: Reproducible scaling laws for contrastive language-image learning. *arXiv preprint arXiv:2309.16671*, 2023.
- [27] Xi Chen et al. Pali-x: On scaling up a multilingual vision and language model. *arXiv preprint arXiv:2407.17453*, 2024.
- [28] Shouda Bai, Jinze Bai, and et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- [29] Jing Zhu, Peng Zhang, Wenhai Wang, and et al. Internvl3: Exploring advanced training and test-time adaptation for native multimodal pretraining. *arXiv preprint arXiv:2504.10479*, 2025.
- [30] Wenhai Wang, Yutong Wang, and et al. Internvl3.5: Advancing open-source multimodal models with cascade reinforcement learning. *arXiv preprint arXiv:2508.18265*, 2025.
- [31] Xingyu An, Haotian Li, and et al. Llava-onevision-1.5: Fully open framework for building vision-language models from scratch. *arXiv preprint arXiv:2509.23661*, 2025.
- [32] Qwen Team. Qwen3-vl. <https://github.com/QwenLM/Qwen3-VL>, 2025. Accessed Oct. 2025.
- [33] Qwen Team. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.
- [34] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.
- [35] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.
- [36] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.
- [37] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *EMNLP*, 2021.
- [38] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *ACL*, 2021.
- [39] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Lu Wang, and Weizhu Wang. Lora: Low-rank adaptation of large language models. *arXiv:2106.09685*, 2022.
- [40] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, et al. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *ICLR*, 2017.

- [41] Peng Wang, An Yang, Rui Men, Junyang Lin, et al. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *ICML*, 2022.
- [42] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the VQA matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017.
- [43] Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019.
- [44] Jun Xu, Tao Mei, Jianfeng Yao, and Yong Rui. MSRVTT-QA: Open-ended video question answering with large-scale datasets. In *ACMMM*, 2017.
- [45] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. NExT-QA: Reasoning on temporal actions and causal effects in video question answering. In *CVPR*, 2021.
- [46] Minghao Li, Tengchao Lv, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Kun Yao, Jianfeng Gao, et al. TrOCR: Transformer-based optical character recognition with pre-trained models. In *NeurIPS*, 2021.
- [47] Geewook Kim, Teakgyu Hong, Sangdoo Yun, Kyungmin Cho, Seong Joon Kim, In So Kweon Kim, and Seunghyun Park. Donut: Document understanding transformer without ocr. In *ECCV*, 2022.
- [48] Matthias Minderer, Alexey Gritsenko, Austin Stone, Ibrahim Alabdulmohsin, et al. Simple Open-Vocabulary Object Detection with Vision Transformers. In *ECCV*, 2022.
- [49] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-Vocabulary Object Detection via Vision and Language Knowledge Distillation. In *CVPR*, 2021.
- [50] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, et al. Segment anything. In *ICCV*, 2023.
- [51] Golnaz Ghiasi, Nenad Tomasev, Dina Bashkirova, Mateusz Malinowski, et al. Scaling open-vocabulary image segmentation with image-level labels. In *ECCV*, 2022.
- [52] Neil Houlsby et al. Parameter-efficient transfer learning for nlp. In *ICLR*, 2019.
- [53] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.
- [54] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.
- [55] Qizhe Xie, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. Self-training with noisy student improves imagenet classification. *CVPR*, 2020.
- [56] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv:1503.02531*, 2015.
- [57] Jason Wei, Maarten Bosma, Vincent Zhao, et al. Finetuned language models are zero-shot learners. In *ICLR*, 2022.
- [58] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv:1807.03748*, 2018.
- [59] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015.
- [60] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017.
- [61] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*, 2016.

- [62] M. A. Rahman and Yang Wang. Optimizing intersection-over-union in deep neural networks for image segmentation. *arXiv:1608.01471*, 2016.
- [63] Haoqiang Fan, Hao Su, and Leonidas J. Guibas. A point set generation network for 3d object reconstruction from a single image. In *CVPR*, 2017.
- [64] Tsung-Yi Lin et al. Microsoft coco: Common objects in context. In *ECCV*, 2014.
- [65] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. In *arXiv:1504.00325*, 2015.
- [66] Jia Deng, Wei Dong, et al. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.
- [67] Christoph Schuhmann, Romain Vencu, Radu Beaumont, Richard Gordon, Ross Wightman, Mehdi Cherti, Clayton Coombes, Srinivasan Katta, Theo Mullis, Mitchell Wortsman, Ludwig Schmidt, Colin Raffel, and Douwe Kiela. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. In *NeurIPS Datasets and Benchmarks*, 2021.
- [68] Christoph Schuhmann, Romain Beaumont, Romain Vencu, Richard Gordon, Ross Wightman, Mehdi Cherti, Clayton Coombes, Srinivasan Katta, Theo Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS Datasets and Benchmarks*, 2022.
- [69] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. In *ECCV*, 2016.
- [70] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016.
- [71] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. In *TACL*, 2014.
- [72] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018.
- [73] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021.
- [74] Kenneth Marino et al. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*, 2019.
- [75] Max Bain, Arsha Nagrani, GÜl Varol, and Andrew Zisserman. Frozen in Time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 2021.
- [76] Jun Xu et al. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016.
- [77] David Chen and William Dolan. Collecting highly parallel data for paraphrase evaluation. In *ACL Workshop*, 2011. MSVD video dataset.
- [78] Luowei Zhou, Chenliang Xu, and Jason J. Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*, 2018.
- [79] Antoine Miech, Dmitry Zhukov, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, 2019.
- [80] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. In *IJCV*, 2017.

- [81] Rohit Girdhar, Kalyan Vasudev Alwala, Mannat Singh, Jeremy Adcock, Mathilde Caron, Ishan Misra, Armand Joulin, Piotr Bojanowski, Alaaeldin El-Nouby, Alexander Kolesnikov, Lucas Beyer, Aravindh Mahendran, Shubham Tulsiani, Chen Sun, Mateusz Malinowski, Gabriel Synnaeve, Andrea Vedaldi, Jack Lanchantin, et al. Imagebind: One embedding space to bind them all. *arXiv:2305.05665*, 2023.